<Bioinformatics Center>Bio-knowledge Engineering

Bioinformatics Center
– Bio-knowledge Engineering –

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
KARASUYAMA, Masayuki
(D Eng)

Assist Prof
NGUYEN, Hao Canh
(Ph D)

Proj Res
NATSUME, Yayoi
(D Agr)

Students

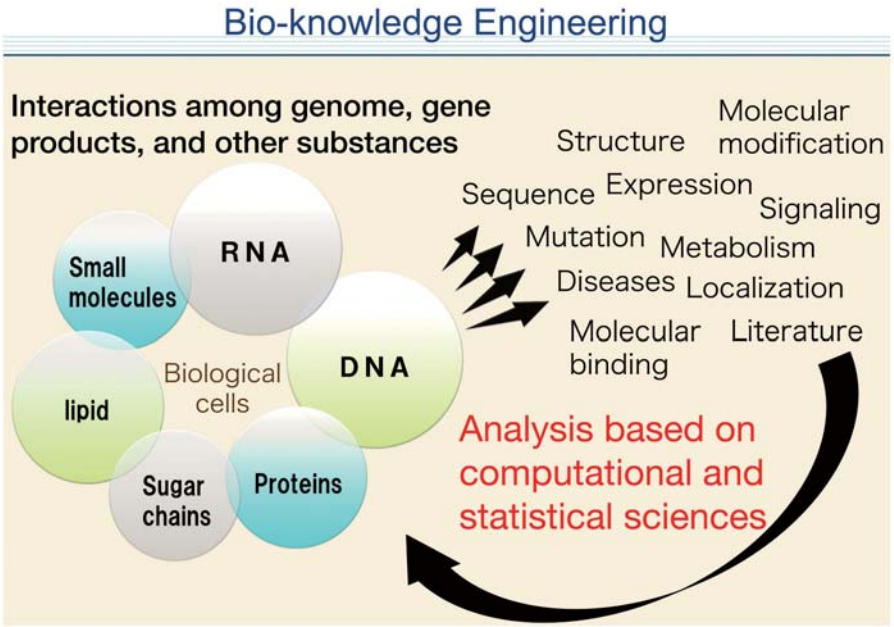
MOHAMED, Ahmed Mohmed (D3) YAMAGUCHI, Shigeru (D2) YOTSUKURA, Sohiya (D2)

Scope of Research

We are interested in graphs and networks in biology, chemistry and medical sciences, which include metabolic networks, protein-protein interactions and chemical compounds. We have developed original techniques in machine learning and data mining for analyzing these graphs and networks, occasionally combining with table-format datasets, such as gene expression and chemical properties. We have applied the developed techniques to real data to demonstrate the performance of the methods and further to find new scientific insights.

KEYWORDS

Bioinformatics
Computational Genomics
Data Mining
Machine Learning
Systems Biology



Selected Publications

Karasuyama, M.; Mamitsuka, H., Manifold-based Similarity Adaptation for Label Propagation, *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS 2013)*, 1547-1555 (2013).

Karasuyama, M.; Mamitsuka, H., Multiple Graph Label Propagation by Sparse Integration, *IEEE Transactions on Neural Networks and Learning Systems*, **24 (12),** 1999-2012 (2013).

Nguyen, C. H.; Wicker, N.; Mamitsuka, H., Selecting Graph Cut Solutions via Global Graph Similarity, *IEEE Transactions on Neural Networks and Learning Systems*, **25 (7),** 1407-1412 (2014).

Mohamed, A.; Hancock, T.; Nguyen, C. H.; Mamitsuka, H., NetPathMiner:R/Bioconductor Package for Network Path Mining through Gene Expression, *Bioinformatics*, **30 (21),** 3139-3141 (2014).

Sparse Multiple Graph Integration for Label Propagation

Predicting labels on a network is an important topic in systems biology and other fields of structured data analysis. For example, the connectivity structure of protein-protein interaction networks (Figure 1) is informative for function category estimation of proteins. A common approach is to propagate information of known function categories to other proteins, of which function categories are unknown, through network connections. In machine learning, this is a problem called *graph-based semi-supervised learning*, because a network can be represented as a "graph" mathematically, and the propagation approach is called "label propagation".

The usefulness of the label propagation algorithms has been demonstrated so far, but their performance highly depends on the way of generating the input graph. For example, in the protein function prediction, various information sources are available such as gene expression, gene sequences and subcellular localization, which can be all given as graphs. We however cannot see the most important graph for prediction. We propose a new approach for the issue of integrating multiple graphs under the label propagation framework.

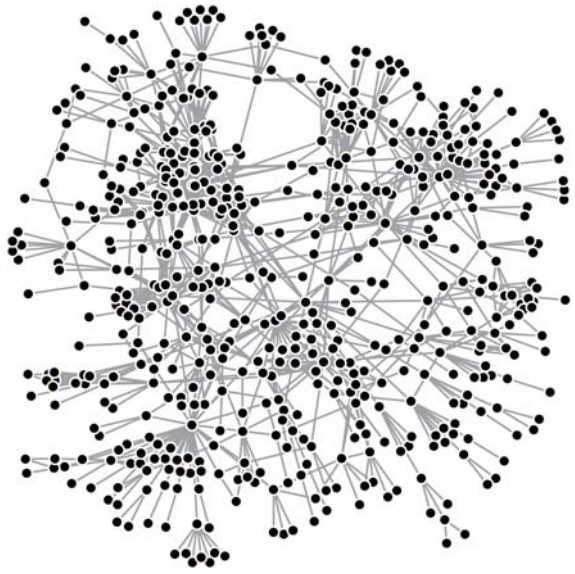


Figure 1. The example of a protein-protein interaction network. Each node corresponds to a protein, and connected protein pairs interact each other.

As already done by the most existing methods, our approach also combines multiple graphs linearly and estimates their weight coefficients. However, our unique property is the sparsity of graph weights. That is, graph weights of our approach can be sparse, meaning that only a part of weights has non-zero values and the rest are equal to exactly 0. This important property provides two advantages. The first advantage is that eliminating irrelevant or noisy graphs in integrating multiple graphs can improve the classification performance. Conventional approaches however have cases of assigning non-zero weights to graphs which are irrelevant to classification (we define such graphs as irrelevant graphs), by which prediction performance can be deteriorated, since irrelevant graphs are kept as the input. On the other hand, our sparseness property allows us to eliminate irrelevant graphs completely because their weights are estimated at zero. The second advantage is that sparse weight coefficients allow to identify the graphs which are important (or not needed) for classification easily. Furthermore, our formulation can provide a clear interpretation of the mechanism of sparsity, and it also offers a kind of grouping effect, which is similar to that given by the standard sparse statistical model called elastic net.

We verified the effectiveness of our approach through synthetic and real-world datasets compared to some other existing approaches (e.g., Figure 2).

Reference

[1] Karasuyama, M.; Mamitsuka, H., Multiple Graph Label Propagation by Sparse Integration, *IEEE Transactions on Neural Networks and Learning Systems*, **24(11),** 1999-2012 (2013).

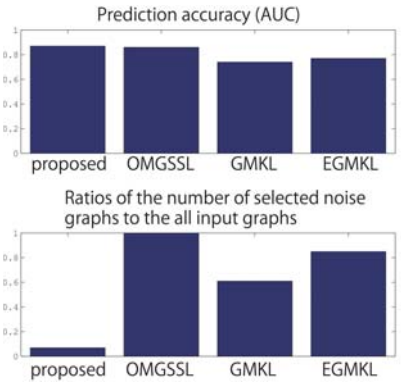


Figure 2. Performance comparison on a protein function prediction. (Top) Prediction accuracy (AUC) of our proposed method and other existing methods. We see that our approach has the highest AUC value in this case. (Bottom) Ratios of the number of selected noise graphs to the all input graphs. In this experiment, we added artificial noise graphs as input graphs. Our approach appropriately removes those noise graphs compared to the other methods.